Michael Saxon

Siegel Fellow
Postdoctoral Scholar
University of Washington
Tech Policy Lab

michael@saxon.me
https://saxon.me

Education

- **Ph.D., Computer Science**
  University of California, Santa Barbara, 2020 – 2025
  Advisor: William Yang Wang.
- **M.S., Computer Engineering**
  Arizona State University, 2018 – 2020
  Advisors: Visar Berisha, Sethuraman Panchanathan.
- **B.S.E., Electrical Engineering**
  Minor, Computational Mathematics
  Arizona State University, 2014 – 2018

Research Interests

- Natural Language Processing (Language models, benchmarks, reasoning, commonsense)
- Multimodal AI (Vision-language models, text-to-image models, video/audio/speech generation)
- Responsible AI (Multilingual & multicultural AI, fairness & transparency, evaluation, safety)

Select Preprints

1. **Michael Saxon***, Xiao Pu*, Wenyue Hua, William Yang Wang, "ThoughtTerminator: Benchmarking, Calibrating, and Mitigating Overthinking in Reasoning Models", preprint, arXiv:2504.13367.
2. Arnav Yayavaram, Siddharth Yayavaram, Simran Khanuja, **Michael Saxon**, Graham Neubig, "CAIRe: Cultural Attribution of Images by Retrieval-Augmented Evaluation", preprint, arXiv:2506.09109.

Peer-reviewed papers

1. Justin Hyundong Cho, Spencer Lin, Tejas Srinivasan, **Michael Saxon**, Deuksin Kwon, Natali T. Chavez, Jonathan May, “Can Vision Language Models Understand Mimes?” To appear in the Findings of the Association for Computational Linguistics, 2025.
2. Weixi Feng, Jiachen Li, **Michael Saxon**, Tsu-jui Fu, Wenhu Chen, William Yang Wang, “TC-Bench: Benchmarking Temporal Compositionality in Text-to-Video and Image-to-Video Generation,” To appear in Findings of the Association for Computational Linguistics, 2025.

3. **Michael Saxon***, Mahsa Khoshnoodi*, Fatima Jahara*, Yujie Lu, Aditya Sharma, William Yang Wang Wang, "Who Evaluates the Evaluations? Assessing the Faithfulness and Consistency of Text-to-Image Evaluation Metrics with T2IScoreScore," **Proc. Conference on Neural Information Processing Systems (NeurIPS) 2024**, *Spotlight, Top 2.5% of 15,000 submissions*, Dec 2024.

4. **Michael Saxon***, Aditya Sharma*, William Yang Wang, "Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts,", **Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024**.

5. **Michael Saxon**, Ari Holtzman, Peter West, William Yang Wang, Naomi Saphra, "Benchmarks as Microscopes: A Call for Model Metrology," **First Conference on Language Modeling (COLM) 2024**

6. **Michael Saxon***, Yiran Luo*, Sharon Levy, Chitta Baral, Yezhou Yang, William Yang Wang, "Lost in Translation? Translation Errors and Challenges for Fair Assessment of Text-to-Image Models on Multilingual Concepts," **Proc. North American Chapter of the Association for Computational Linguistics 2024**, *Oral (5% of subs.)* June 2024.

7. Liangming Pan, **Michael Saxon**, Wenda Xu, Deepak Nathani, Xinyi Wang, William Yang Wang, "Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies," **Trans. of the Association for Computational Linguistics (TACL)** May 2024.

8. Vaishnavi Himakunthala*, Andy Ouyang*, Daniel Rose*, Ryan He*, Alex Mei, Yujie Lu, C. Sonar, **Michael Saxon**, William Yang Wang, "Let's Think Frame by Frame with VIP: A Video Infilling and Prediction Dataset for Evaluating Video Chain-of-Thought," **Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023**, Dec 2023

9. Xinyi Wang, Wanrong Zhu, **Michael Saxon**, Mark Steyvers, William Yang Wang, "Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning," **Proc. Conference on Neural Information Processing Systems (NeurIPS) 2023**, Dec 2023

10. **Michael Saxon**, William Yang Wang, "Multilingual Conceptual Coverage in Text-to-Image Models," **Proc. Association for Computational Linguistics; FAccT 2023 Oral** Jul 2023.

11. Yi-lin Tuan, Alon Albalak, Wenda Xu, **Michael Saxon**, Connor Pryor, Lise Getoor, William Yang Wang, "CausalDialogue: Modeling Utterance-level Causality in Conversations," **Findings of the Association for Computational Linguistics**, Jul 2023.

12. Matthew Ho, Aditya Sharma, Justin Chang, **Michael Saxon**, Sharon Levy, Yujie Lu, William Yang Wang, "WikiWhy: Answering and Explaining Cause-and-Effect Questions", **Proc. International Conference on Learning Representations 2023**, Oral Paper: *Top 5% out of all 4019 submissions*, Kigali, Rwanda, May 1st to 5th.

13. Xinyi Wang, **Michael Saxon**, Jiachen Li, Hongyang Zhang, Kun Zhang, William Yang Wang, "Causal Balancing for Domain Generalization", **Proc. International Conference on Learning Representations 2023**, Kigali, Rwanda, May 1st to 5th.

14. **Michael Saxon**, Xinyi Wang, Wenda Xu and William Yang Wang, "PECO: Examining Single Sentence Label Leakage in Natural Language Inference Datasets through Progressive Evaluation of Cluster Outliers", in Proceedings of The 17th Conference of the European Chapter of the Association for Computational Linguistics (**Proc. European Chapter of the Association for Computational Linguistics**), long paper, Dubrovnik, Croatia, May 2-6 2023, ACL.

15. Wenda Xu, Yi-Lin Tuan, Yujie Lu, **Michael Saxon**, Lei Li, William Yang Wang, "Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis", **Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021**, pp 6559–6574, long paper ACL.

16. Wenda Xu, **Michael Saxon**, Misha Sra, William Yang Wang, "Self-Supervised Knowledge Assimilation for Expert-Layman Text Style Transfer", to appear in Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (**AAAI 2022**), long paper, Vancouver, BC, Canada.

17. Xinyi Wang, Wenhu Chen, **Michael Saxon**, William Yang Wang, "Counterfactual Maximum Likelihood Estimation for Training Deep Networks", (**Proc. Conference on Neural Information Processing Systems (NeurIPS) 2021**), long paper, online.

18. **Michael Saxon**, Sharon Levy, Xinyi Wang, Alon Albalak and William Yang Wang, "Modeling Disclosive Transparency in NLP Application Descriptions", to appear in (**Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021**), *Oral (8% of subs.)* long paper, online, ACL.

19. **Michael Saxon**, Samridhi Choudhary, Joe McKenna, Athanasios Mouchtaris, "End-to-End Spoken Language Understanding for Generalized Voice Assistants," **Interspeech 2021**, pp. 4738–4742.

20. Sharon Levy, **Michael Saxon** and William Yang Wang, "Investigating Memorization of Conspiracy Theories in Text Generation", to appear in Findings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**Findings of ACL-IJCNLP 2021**), long paper, online, ACL.

21. **Michael Saxon**, Ayush Tripathi, Yishan Jiao, Julie Liss, Visar Berisha, "Robust Estimation of Hypernasality in Dysarthria," **IEEE Trans. on Audio, Speech, and Language Processing** 2020, Vol. 28, pp. 2511–2522.

22. **Michael Saxon***, Joe McKenna*, Samridhi Choudhary*, Grant Strimel, Athanasios Mouchtaris, "Semantic Complexity in End-to-End Spoken Language Understanding," **Interspeech 2020**, pp. 4273–4277.

23. Meredith Moore, P. Papreja, **Michael Saxon**, Visar Berisha, Sethuraman Panchanathan, "UncommonVoice: A Crowdsourced Dataset of Dysphonic Speech," **Interspeech 2020**, pp. 2532–2536.

24. Meredith Moore, **Michael Saxon**, Hemanth Venkateswara, Visar Berisha, Sethuraman Panchanathan, "Say what? A dataset for exploring the error patterns that two ASR engines make," **Interspeech 2019**, pp. 2528–2532.

25. **Michael Saxon**, Julie Liss, Visar Berisha, "Objective Measures of Plosive Nasalization in Hypernasal Speech," 2019 **IEEE ICASSP 2019**, pp. 6520–6524.

26. Todd Houghton, **Michael Saxon**, Zeming Song, Hoa Nyugen, Hanqing Jiang and Hongbin Yu, "2D Grating Pitch Mapping of a through Silicon Via (TSV) and Solder Ball Interconnect Region Using Laser Diffraction" **IEEE 66th Electronic Components and Technology Conference (ECTC) 2016**, pp. 2222–2227. *(Texas Instruments Best Student Interactive Paper Award)*

Peer-reviewed workshop, demo & non-archival presentations

1. Avani Tanna, **Michael Saxon**, Amr El Abbadi, William Yang Wang, "Data Augmentation for Diverse Voice Conversion in Noisy Environments," **Interspeech 2023 Show and Tell**, Aug 2023.

2. **Michael Saxon**, William Yang Wang, "Disparities in Text-to-Image Model Concept Possession Across Languages," **FAccT 2023 Oral** (Non-archival), Jun 2023.

3. **Michael Saxon***, Samarth Bhandari*, Lewis Ruskin, Gabrielle Honda, "Word Pair Convolutional Model for Happy Moment Classification," **AAAI AffCon Workshop 2019**, pp. 111–119. *(Oral; Shared task 2nd place)*

Other preprints and non-archival works

1. Qiucheng Wu, Handong Zhao, **Michael Saxon**, Trung Bui, William Yang Wang, Yang Zhang, Shiyu Chang, "VSP: Assessing the dual challenges of perception and reasoning in spatial planning tasks for VLMs," preprint, arXiv:2407.01763.
2. Vaishnavi Himakunthala*, Andy Ouyang*, Daniel Rose*, Ryan He*, Alex Mei, Yujie Lu, Chinmay Sonar, **Michael Saxon**, William Yang Wang, "Visual Chain of Thought: Bridging Logical Gaps with Multimodal Infillings," *preprint*, arXiv:2305.02317.
3. Alex Mei, **Michael Saxon**, Shiyu Chang, Zachary C Lipton, William Yang Wang, "Users are the North Star for AI Transparency," *preprint*, arXiv:2303.05500, Mar 2023.

Mentees

1. Siddharth Yayavaram (BITS Pilani/CMU 2024)
2. Arnav Yayavaram (BITS Pilani/CMU 2024)
3. Aditya Sharma (UCSB BS/MS 2024)
4. Mahsa Khoshnoodi (Fatima Fellowship 2023; now Ph.D. Georgetown University)
5. Fatima Jahara (Fatima Fellowship 2023; now Ph.D. Rutgers University)
6. Eshaan Tanwar (IIT Delhi 2023)
7. Avani Tanna (UCSB MS 2023)
8. Andy Ouyang (UCSB ERSP 2022)
9. Ryan He (UCSB ERSP 2022)
10. Vaishnavi Himakunthala (UCSB ERSP 2022)
11. Daniel Rose (UCSB ERSP 2022)
12. Matthew Ho (UCSB ERSP 2021, now Ph.D.; University of California, San Diego)
13. Aditya Sharma (UCSB ERSP 2021)
14. Justin Chang (UCSB ERSP 2021)

Press coverage

- "AI Models Embrace Humanlike Reasoning: Researchers are pushing beyond chain-of-thought prompting to new cognitive techniques." Edd Gent, IEEE Spectrum, 08 May 2025.
- "Gemini's data-analyzing abilities aren't as good as Google claims." Kyle Wiggers, TechCrunch, June 29 2024.
- "Groundswell of Opposition to CA's AI Bill as it Nears Vote." Brandon Gorrell and Riley Nork, PirateWires, Aug 13, 2024

Awards

- **Rising Star in Generative AI** *1/9 selectees, UMass Amherst Rising Stars Workshop* 2024
- **Google PhD Fellowship Nominee** *One of 4 selected by UCSB* 2024
- **Neal Fenzi—Resonant Founder Fellowship** *University of California, Santa Barbara* 2024
- **Association for Computational Linguistics** *Oustanding Reviewer Award* 2023
- **National Science Foundation Graduate Research Fellowship** *(NSF GRFP)* 2020
- **Center for Responsible Machine Learning Fellowship** *University of California, Santa Barbara* 2020
- **Graduate Division Central Fellowship** *University of California, Santa Barbara* 2020
- **Phi Kappa Phi** *Inductee* 2016
- **IEEE Eta Kappa Nu** (HKN) *Inductee* 2015
- **Presidential Scholarship** (Full Tuition) *Arizona State University* 2014

Presentations & Invited talks

- **Google Translate Research**, *Seminar talk* 6/2025
- **How to nitpick multimodal evaluations**, *CVPR 2025 Tutorial* 6/2025
- **Stanford University Choi-Lab**, *Invited talk* 1/2025
- **Graduate Seminar Series**, *Lockheed Martin Santa Barbara Focalplane* 10/2024
- **Rising Stars in Generative AI Workshop**, *University of Massachusetts, Amherst* 9/2024
- **Allen Institute for AI**, *Company Talk* 9/2024
- **Stanford University** *SALT Group Presentation* 8/2024
- **University of Maryland, College Park** *UMD CLIP Seminar* 5/2024
- **Georgetown University** *NLP Group Presentation* 5/2024
- **University of Maryland, Baltimore** *Perception, Prediction, and Reasoning Seminar* 4/2024
- **Arizona State University** *Active Perception Group Presentation* 11/2023
- **USC Information Sciences Institute** *Natural Language Processing Seminar* 11/2023

Service

- *Program Co-Chair*, 2022 Southern California NLP Workshop (SoCalNLP) 11/2022
- *Reviewer*, AAAI, EMNLP, ACL, EACL, ARR, NeurIPS, ICLR, FAccT, ICASSP, Interspeech 2020–*present*